PB82-132499

Fourier Encoded Data Searching of Infrared Spectra (FEDS/IRS)

(U.S.) Environmental Research Lab.
Athens, GA

24 Mar 81

U.S. Department of Commerce
National Technical Information Service
NTIS.

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/J-81-144	2. Journal Article	3. RE P882-132499
4. TITLE AND SUBTITLE Fourier Encoded Data Searching of Infrared Spectra (FEDS/IRS)		5. REPORT DATE November 1981 6. PERFORMING ORGANIZATION CODE 4337066
7. AUTHOR(S) L.V. Azarraga*, R.R. Williams** and J.A. deHaseth**		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS *Environmental Research Laboratory Athens, Georgia 30613 **University of Alabama Tuscaloosa, Alabama 35486		10. PROGRAM ELEMENT NO. CBNC1A 11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS U.S. Environmental Protection Agency--Athens GA Environmental Research Laboratory College Station Road Athens, Georgia 30613		13. TYPE OF REPORT AND PERIOD COVERED Journal Article 14. SPONSORING AGENCY CODE EPA/600/01

15. SUPPLEMENTARY NOTES
Published in Journal of Applied Spectroscopy. 35(5):466-469,1981.

Contact: L.V. Azarraga, FTS 250-3552

16. ABSTRACT

A new library searching technique is reported that relies on Fourier transforms of infrared (IR) absorbance spectra. Searching in the time domain is shown to be more tolerant to noise than searches in the spectral domain and fewer points are required to encode the unique characteristics of the individual spectrum. This decreases both search time and library storage requirements. Additionally, time-domain classification of IR spectra is demonstrated and the possibility of indexing the library for a binary search is discussed.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS (This Report) UNCLASSIFIED	21. NO. OF PAGES 4
	20. SECURITY CLASS (This page) UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (9-73)

Fourier Encoded Data Searching of Infrared Spectra (FEDS/IRS)

L. V. AZARRAGA,* R. R. WILLIAMS, and J. A. DE HASETH†
Environmental Research Laboratory, U.S. Environmental Protection Agency, Athens, Georgia 30613

A new library searching technique is reported that relies on Fourier transforms of infrared (IR) absorbance spectra. Searching in the time domain is shown to be more tolerant to noise than searches in the spectral domain and fewer points are required to encode the unique characteristics of the individual spectrum. This decreases both search time and library storage requirements. Additionally, time-domain classification of IR spectra is demonstrated and the possibility of indexing the library for a binary search is discussed.

Index Headings: **Computer, applications; Infrared; Infrared search system.**

INTRODUCTION

Although programs for interpretation of infrared spectra are in the developmental stage, spectral library searching is the current method of choice in commercially available Fourier transform infrared (FT-IR) spectrometers. Unknown spectra are compared point-by-point with library entries and an Euclidean distance is calculated for each. Spectra are ranked according to distance and the smallest distances are reported as "hits." This method is extremely reliable when the unknown spectrum is in the library. However, it suffers from three serious drawbacks. (1) The search is slow and more importantly the time required increases as the size of the library increases. For example, the spectral search algorithm[1] of the Gas Chromatography-Infrared Fourier Transform Software (GIFTS)[2] developed in our laboratory can search its 3300 entry library in approximately 5 min. Because this library is constantly updated and expanded, the search will ultimately require excessive amounts of time. (2) The storage space allocated for the library also increases as library size is expanded, thus increasing the required memory size of the computer and the access times in reading the library. One way to minimize this problem is to reduce the resolution of the spectrum or compress the spectral information into as small a volume as possible (without losing resolution). (3) Finally, missing data points in the spectra grossly affect the quality of the search.

We have developed a Fourier Encoded Data Search (FEDS) algorithm for infrared searching that performs the point-by-point comparison in the time domain rather than in the frequency domain. In view of Fellgett's advantage,[3] each datum in the time domain contains information on all frequencies present in the spectrum. Therefore, fewer points are required to define the signature of each compound and frequency information is not completely lost when points are missing. Absorbance spectra of unknowns are zero filled leaving only the fingerprint region, which is Fourier transformed. This is done simply to limit the spectrum to the region in which the MCT detector, which is commonly used in GC/FT-IR, has its optimum response. Otherwise, this step is not necessary. If a detector for GC/FT-IR has sufficient sensitivity throughout the mid-IR region, the entire spectrum may be Fourier transformed without affecting the size of the data used for comparison. This could further enhance the accuracy of the search system.

I. EXPERIMENTAL

A. Hardware. A Digilab FTS-14C Fourier transform spectrometer with a mercury-cadmium-telluride detector was used to acquire sample interferograms from a Varian 3700 gas chromatograph, equipped with capillary GC columns. The instrument was controlled by a Data General Nova 3 computer with 64K of MOS memory using the GIFTS program.

B. Software. A library of interferograms was generated by condensing the 4 cm^{-1} resolution vapor phase IR spectra from EPA library (EPALIB) to give 8 cm^{-1} resolution spectra, which are then zero filled, leaving only the 700 to 2000 cm^{-1} region and finally transformed to give 512-point interferograms. All programming was done in FORTRAN except when speed was required; assembler language routines were used in these special cases.

II. RESULTS AND DISCUSSION

A. Standard Library Searching. In the FEDS technique, an interferogram is represented by a 60-dimensional vector and comparisons are made by taking vector dot products between sample and library entries. When two vectors are identical, they fall exactly on top of one another and so have a dot product of 1.0. Theoretically these dot products may run from −1.0 to 1.0. If this process were carried in the spectral domain, however, values could only range from 0.0 to 1.0 because all spectral data points represent absorbance intensities and are therefore positive. By using the bipolar interferogram, the FEDS algorithm has twice the dynamic range of an equivalent search performed in the spectral domain.

To evaluate the performance of FEDS/IRS, individual absorbance spectra were calculated for each of the interferograms across a sample effluent peak from the GC. The spectra represent the same compound at various

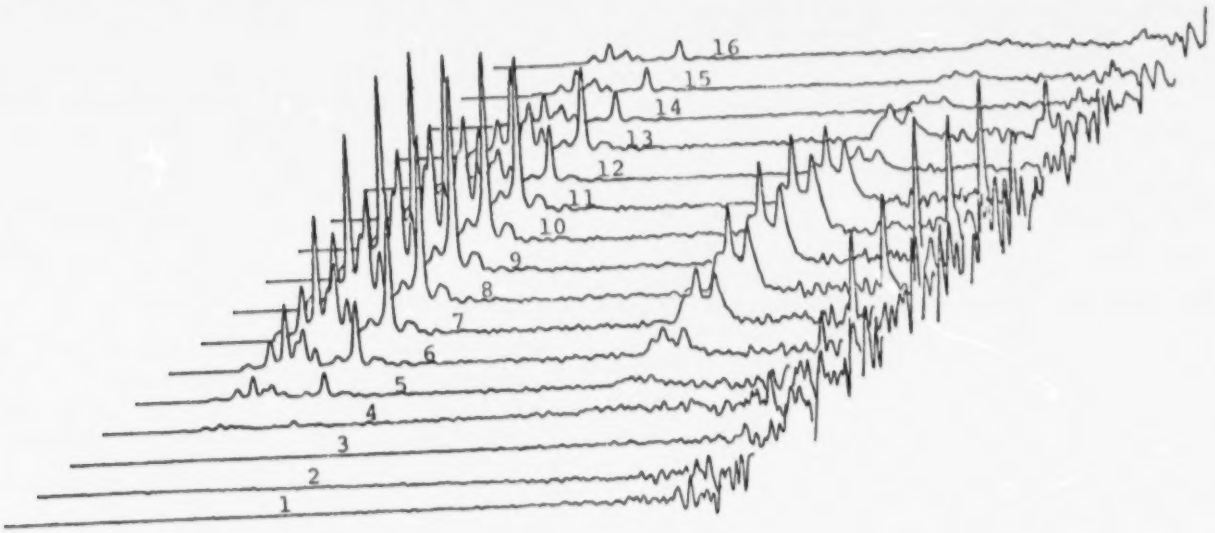
Received 24 March 1981.
* Author to whom correspondence should be addressed.
† Present address: Department of Chemistry, University of Alabama, Tuscaloosa, AL 35486.
Mention of trade names or commerical products does not constitute endorsement or recommendation for use by the U.S. Environmental Protection Agency.

signal to noise ratios (SNR). Both GIFTS and FEDS/IRS were used to identify each of the spectra.

Fig. 1 shows the series of absorbance spectra computed from successive interferograms taken as a sample of 2,4-xylenol eluted from the GC. The lower portion of Fig. 1 contains the search results for each spectrum as reported by GIFTS and FEDS. It is seen that both methods give correct results for spectra 5 through 16. In the case of spectrum 4, GIFTS reports carbon monoxide, whereas FEDS reports 4-t-butyl-o-cresol, a compound whose spectrum between 700 to 2000 cm^{-1} is very similar to 2,4-xylenol. In fact, the FEDS method did report 2,4-xylenol as the second hit.

B. Library Indexing. The concept of ordering the library with respect to an arbitrarily selected vector representing a compound in the library raises the possibility of generating a binary searching algorithm. In a library with this characteristic, it would be possible to search through great numbers of entries in very few steps. For instance, a 32 000 entry library could be searched in no more than 16 comparisons. Further, the size of the library can be doubled and only one additional comparison is necessary for a complete search.

To investigate this possibility, we chose various vectors from the library and computed dot products for all other library members ranking them with respect to the value of their dot products. Geometrically, we have simply measured the projection of each library vector onto the



	GIFTS	FEDS/IRS
1.	AZOBENZENE, 2,2*-DICHLORO-,	PHENYLETHYLALCOHOL,o-NITRO,
2.	CARBON MONOXIDE	ANILINE, 2,5-DICHLORO,
3.	CARBON MONOXIDE	BENZYLAMINE, p-FLUORO,
4.	CARBON MONOXIDE	o-CRESOL, 4-t-BUTYL (SEE NOTE)
5.	2,4-XYLENOL	2,4-XYLENOL
6.	2,4-XYLENOL	2,4-XYLENOL
7.	2,4-XYLENOL	2,4-XYLENOL
8.	2,4-XYLENOL	2,4-XYLENOL
9.	2,4-XYLENOL	2,4-XYLENOL
10.	2,4-XYLENOL	2,4-XYLENOL
11.	2,4-XYLENOL	2,4-XYLENOL
12.	2,4-XYLENOL	2,4-XYLENOL
13.	2,4-XYLENOL	2,4-XYLENOL
14.	2,4-XYLENOL	2,4-XYLENOL
15.	2,4-XYLENOL	2,4-XYLENOL
16.	2,4-XYLENOL	2,4-XYLENOL

NOTE:
IN SPECTRUM #4 FEDS IDENTIFIED 2,4-XYLENOL AS THE SECOND HIT.

FIG. 1. Three-dimensional reconstruction of the elution of sample from GC/FT-IR. Identifications by GIFTS and FEDS/IRS are also given.

reference vector. Table I contains the results for an ordering with respect to hexane. The compounds most similar to hexane show up at the top of the list. More interestingly, 1-hexanol and 2-hexanol were used for the listings in Tables II and III. Here the best "hits" are almost exclusively compounds having substitutions of hydroxyl groups at the 1- and 2-positions, respectively. This particular type of indexing is sensitive to both substituent group and substitution pattern.

An attempt to order the library with respect to 1-hexyne resulted in the listing of Table IV. In an apparent error, FEDS/IRS picked out mainly carboxylic acids for matches; Fig. 2 discloses the source of this error. The so-called 1-hexyne spectrum from the vapor phase IR library contains several mysterious peaks, notably those at 3580 cm^{-1} (carboxylic OH stretch region) and 1760 cm^{-1} (C=O stretch region). Apparently the spectrum is that of a mixture of 1-hexyne and hexanoic acid. It should be

TABLE I. FEDS/IRS ordering of 3300 EPALIB spectra using hexane as a reference vector.[a]

Hit no.	Serial no.	DOT PDT	Name
1	39	1.0000	Hexane
2	46	0.9909	Pentane
3	415	0.9987	Decane
4	545	0.9883	Heptane
5	887	0.9870	Nonane
6	1573	0.9828	Pentane, 3-ethyl-,
7	434	0.9827	Hendecane
8	432	0.9827	Dodecane
9	2001	0.9754	Heptanenitrile, 3-ethyl
10	895	0.9752	Octane
11	889	0.9751	Tetradecane
12	84	0.9744	Tridecane
13	327	0.9719	Cyclohexane, 1,4-dimethyl
14	2189	0.9689	Nonanenitrile
15	2206	0.9671	Undecanenitrile
16	470	0.9670	Hexadecane
17	646	0.9640	Cyclohexane, 1,1-dimethyl
18	483	0.9634	Heptadecane
19	1922	0.9618	Cyclohexane, trans-1,2-dimethyl
20	42	0.9599	Pentane, 3-methyl-,

[a] Nitriles appear because the region above 2000 cm^{-1} was not included in transform.

TABLE II. Library ordering by FEDS technique using 1-hexanol as reference.[a]

Hit no.	Serial no.	DOT PDT	Name
1	412	0.9999	1-Hexanol
2	419	0.9989	1-Heptanol
3	4	0.9977	1-Octanol
4	396	0.9964	1-Nonanol
5	437	0.9941	1-Pentanol, 4-methyl-,
6	459	0.9939	1-Decanol
7	380	0.9936	1-Butanol
8	2694	0.9919	Heptanenitrile, 7-hydroxy
9	847	0.9910	Isooctyl alcohol
10	466	0.9907	1,5-Pentanediol
11	438	0.989	1-Pentanol, 3-methyl-
12	235	0.9892	Cyclohexanol, 2-methyl
13	391	0.9886	1-Butanol, 3-methyl-,
14	1882	0.9863	1-Undecanol
15	1581	0.9861	Cyclohexanepropanol
16	346	0.9858	1,5-Pentanediol, 3-methyl
17	1571	0.9852	Decane, 1-fluoro-,
18	1633	0.9836	1-Dodecanol
19	602	0.9832	1,2-Butanediol
20	2020	0.9826	1-Butanol, 4-/dipropylamino/-

[a] Note the predominance of 1-hydroxy alkanes.

TABLE III. Library ordering by FEDS technique using 2-hexanol as reference.[a]

Hit no.	Serial no.	DOT PDT	Name
1	439	1.0000	2-Hexanol
2	612	0.9828	2-Octanol
3	384	0.9786	2-Heptanol
4	1212	0.9789	2-Pentanol, 2-ethyl-,
5	1028	0.9756	2-Hexanol, 3-methyl-,
6	156	0.9729	2-Nonanol
7	389	0.9718	2-Pentanol
8	378	0.9691	2-Pentanol, 4-methyl-,
9	211	0.9673	2-Butanol, 3-methyl-,
10	321	0.9670	2-Decanol
11	2790	0.9595	Cyclo pentanemethanol, A-methyl-,
12	316	0.9565	2-Heptanol, 5-ethyl-,
13	1016	0.9560	2,5-Hexanediol
14	1233	0.9538	2,4-Pentanediol, 3-methyl
15	1931	0.9480	2-Nonanol, 5-ethyl-,
16	320	0.9384	Cyclohexanemethanol
17	1565	0.9238	2-Tridecanol
18	1566	0.9180	2-Butanol, 3,3-dimethyl
19	2021	0.1943	2,4-Nonanediol, 5-ethyl-3-methyl-
20	554	0.9113	4-Heptanol, 2,6-dimetyl

[a] Note the predominance of 2-hydroxy substituted alkanes as compared to Table II.

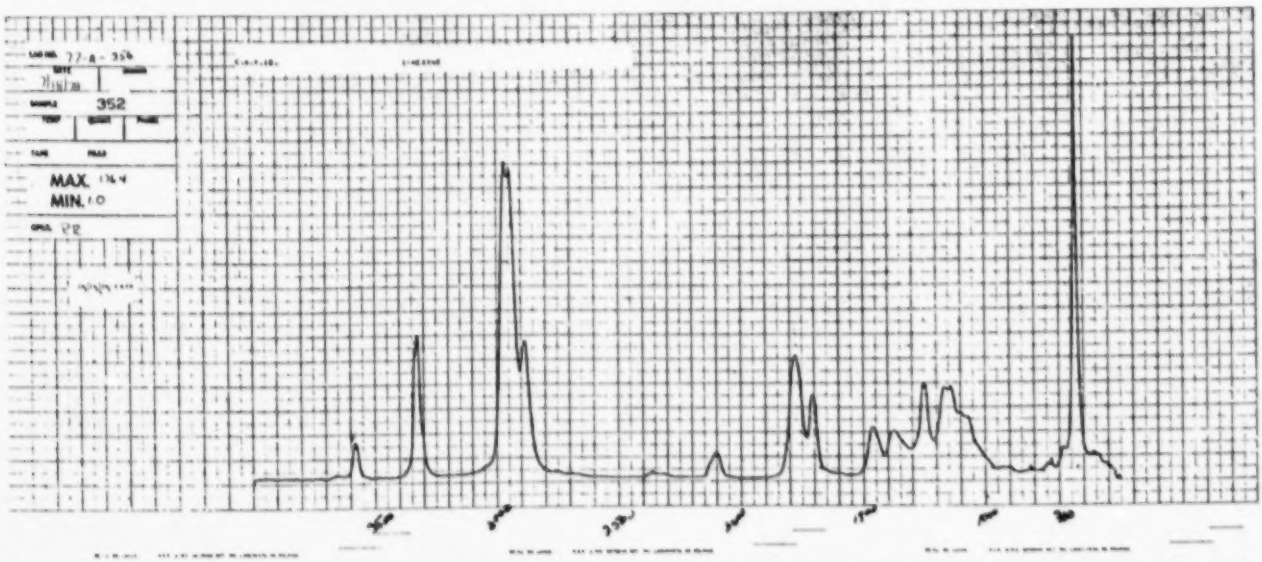


FIG. 2. Library spectrum of 1-hexyne used in ordering for Table IV. Note carboxyl OH stretch at 3580 cm^{-1} and carbonyl stretch at 1780 cm^{-1}. Spectrum obviously contains impurities that are probably oxidized 1-hexyne.

TABLE IV. Library ordering using 1-hexyne as reference.[a]

Hit no.	Serial no.	DOT PDT	Name
1	352	1.0000	1-Hexyne
2	2048	0.9482	Suberic acid
3	1412	0.9436	Azelaic acid
4	1144	0.9367	Arachidic acid
5	3263	0.9359	Butyric acid, 4-acetyl-,
6	1456	0.9326	Decanedioic acid
7	3071	0.9311	Propionic acid, 3,3-diphenyl-,
8	2217	0.9294	Glutaric acid, methyl ester
9	2067	0.9270	Glutaric acid, 2-methyl
10	292	0.9251	Levulinic acid
11	349	0.9239	Azelaic acid, monomethyl ester
12	2024	0.9234	Levulinic acid
13	2427	0.9208	Heptadecanoic acid
14	2095	0.9135	2-Thiophenebutyric acid
15	1160	0.9133	Myristic acid
16	2081	0.9130	2-Dodecanoic acid
17	195	0.9127	Oleic acid
18	1189	0.9125	Behenic acid
19	3161	0.9116	Proprionic acid
20	1727	0.9108	Butyric acid, 3-bromo

[a] Ordering appears to be faulty. See text for explanation.

noted here that the region above 2000 cm^{-1} was not transformed; the FEDS system had only the fingerprint region for comparison.

III. CONCLUSIONS

The use of time domain data in a library search has been successfully demonstrated. The method proves to be both more reliable and more compact than currently available techniques. Moreover, and perhaps more importantly, the vector representation of interferogram manipulation has proven effective in classifying infrared spectra of organic compounds.

Work is under way in our laboratory towards implementing a truly binary search algorithm using these vector manipulations.

1. A. Hanna, J. C. Marshall, and T. L. Isenhour, J. Chromatographic Science, **17**, 434 (1979).
2. L. V. Azarraga and D. A. Hanna, *GIFTS, Athens ERL GC/FT-IR Software and Users' Guide* (U.S. EPA/ERL, Athens, GA, 1979).
3. P. Fellgett, J. Phys. Radium, **19**, 187 (1958).